6M-05

マルチモーダル情報を利用した家庭用ロボットのためのロバストな音声命令理解

小堀　嵩博 †　中村　友昭 †　長井　隆行 †
岩橋　直人 ‡　船越　孝太郎 §　金子　正秀 †
† 電気通信大学　‡ 岡山県立大学　§(株) ホンダ・リサーチ・インスティチュート・ジャパン

1　はじめに

本稿では，RoboCup@Home[*1]において実施されている GPSR (General Purpose Service Robot) タスクを対象とし，ロバストな音声命令理解モデルを提案する．RoboCup@Home とは家庭用ロボットの性能の向上を目的とした競技会であり，GPSR タスクはその中の 1 つのタスクとなっている．GPSR タスクでは，ユーザが音声によりロボットへ命令をし，その命令の理解度と命令の達成度が評価される．また，このタスクでは，どのような言い回しで命令が提示されるか明確に決まっていないため，ロボットは命令を柔軟に理解できる必要がある．さらに，競技の解説者の声や，観客の声などのノイズの影響により，音声の誤認識も発生する．このような問題を解決するため，我々は物体の存在確率と言語情報を統合した音声命令理解モデルを提案した [1]．しかし，[1] のモデルでは，物体の存在確率の取得を実環境に近い条件で行っておらず，環境内の物体が移動してしまうような場合に対処することが困難であった．そこで，本稿では，GMM (Gaussian Mixture Model) を用いて物体の存在確率をモデル化し，より実環境に近い条件でも対応できる音声命令理解モデルを提案する．

GPSR タスクの音声命令理解には構文解析を利用する手法が多く提案されている [2, 3]．しかし，構文解析のみを用いた手法では音声の誤認識が発生すると解析が困難となる．一方，我々のモデルは，物体の存在確率と言語情報を統合することで音声の誤認識に対しても頑健である．また，周囲の環境情報を用いることで音声認識の言語モデルを変更し，音声の認識精度を向上させる方法が提案されている [4]．しかし，この手法は周囲の積み木の状況を利用して言語モデルの変更を行っており，GPSR タスクのような様々な物体が存在する複雑な環境下で有効であるかは明らかではない．

2　音声命令理解モデルの概要

ここでは，我々が提案する音声命令理解モデルについて概要を述べる．モデルの詳細は先行研究 [1] を参照されたい．本稿では，この音声命令理解モデルの物体の存在確率を改善する．図 1 が提案する音声命令理解モデルのグラフィカルモデルである．まずユーザ発話 u を音声認識し，その文字列 r を取得する．その文字列 r に対し，ユーザの命令意図を推定する．本研究では，ユーザの命令意図を行動タイプ a とスロット s で定義している．行動タイプはロボットがすべき行動であり，スロットはその行動を実行するために必要な情報である．例えば，「リビングに向かって」というような命令では，行動タイプは “go”，スロットはロボットが行くべき場所である “リビング” となる．ここで，スロットを推定するときに文字列 r のみではなく，行動タイプとスロットの共起関係 $P(s|a)$ と環境中にそのスロットが存在する確率 $P(s|e)$ を利用する．この情報を付与することで，より高い精度で命令を理解をすることが可能である．

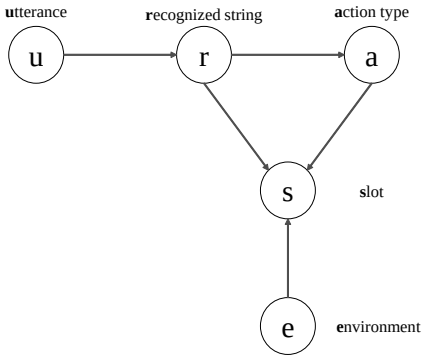

図 1　音声命令理解のグラフィカルモデル

本研究では，行動タイプの識別には Support Vector Machine (SVM) を用い，スロット抽出には Conditional Random Fields (CRF) を用いた．さらに，行動タイプとスロットの共起関係 $P(s|a)$ は，行動タイプに必要なスロットに対する抽出されたスロットの F 値を利用した．$P(s|e)$ はロボットが環境から取得できる情報の一つである物体の存在確率とし，そのスコアを利用した．

3　物体の存在確率

より複雑な環境下で音声命令を理解するために，本稿では GMM を用いて物体の存在確率をモデル化する．ロボットは環境内をアクティブ探索法 [5] を用いて物体認識を行い，k 回目の観測において物体 o がある地点 $\bar{\boldsymbol{x}}_k$ に存在する確率 $P_k(o)$ を取得することができる．しかし物体認識では，必ずしも正確な位置を認識できるとは限らず，誤差が生じる可能性がある．そこで，この物体の位置を，$\bar{\boldsymbol{x}}_k$ を平均とした 2 次元のガウス分布 $\mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \boldsymbol{\sigma})$ を用いて表現する．さらに家庭内において，物体は必ずしも同じ場所にあるとは限らない．例えば，本などであ

Robust Spoken Instruction Understanding for a Domestic Service Robot Using Multimodal Information
†Takahiro KOBORI　†Tomoaki NAKAMURA　†Takayuki NAGAI
‡Naoto IWAHASHI　§Kotaro FUNAKOSHI　†Masahide KANEKO
†The University of Electro-Communications
‡Okayama Prefectural University
§Honda Research Institute Japan Co., Ltd.

[*1] RoboCup@Home, http://www.robocupathome.org/.

ればロボットが観測した後に，ユーザが本棚から取り出し，どこかへ持っていく可能性がある．すなわち，存在確率は時間が経過するにつれ，信頼度が低下すると考えられる．そこで，物体 o の k 回目の観測からの経過時間 t_k によって，$P_k(o)$ が $\exp(-\lambda t)$ の割合で減衰すると考える．すなわち，k 回目の観測から，物体 o が場所 $\boldsymbol{x}$ に存在する確率は以下のように表わすことができる．

$$P_k(\boldsymbol{x}|o) \propto \exp(-\lambda t_k) P_k(o) \mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \boldsymbol{\sigma}) \quad (1)$$

ロボットは，環境を探索することで，複数回物体 o を観測することになる．そこで，物体 o を K_o 回観測した場合，式 (1) の確率分布を重ね合わせることで，物体の存在確率を表現する．

$$P(\boldsymbol{x}|o) \propto \sum_{k=1}^{K_o} \exp(-\lambda t_k) P_k(o) \mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \boldsymbol{\sigma}) \quad (2)$$

この式は，$\exp(-\lambda t_k)P_k(o)$ を混合比とした GMM と見なすことができる．物体の存在スコアは式 (2) を正規化し，最も確率が高かった地点の値を用いる．

4 評価実験

提案するモデルが GPSR タスクにおける音声命令理解において有効か検証した．RoboCup@Home の環境に近い環境で音声命令理解を行うために，シミュレータ SIGVerse[*1]によって再現された 2LDK の空間で実験を行った．物体の存在確率は以下のように取得した．

1. ロボットは物体認識をしながら，事前に登録した物体が存在する場所へ移動する
2. ロボットがすべての登録した場所を探索した後，物体の配置を変更する
3. 配置を変更後，ロボットは再び物体認識をしながら移動する

これを 5 回行い，環境情報の取得を行った．物体の配置の変更では，環境内の特定の 2 つの物体をランダムに選択された机へ移動し，他の物体は，2 次元のガウス分布に従い同じ机上で移動させた．本実験では，$\sigma^2\boldsymbol{I}$ とした等方的な分散共分散行列のガウス分布を用い，$3\sigma = 20$[cm] とした．

実験に用いるデータセットは，2011 年の GPSR タスクで使用された命令文を自動生成するソフトウェア[*2]と，研究室に所属する学生にアンケートを実施し収集した命令文 1569 文である．以下が収集した命令文の一例である．

- ウッドテーブルにあるテディベアを掴んで，キッチンにいるエマさんに挨拶して，テディベアを渡して
- テディベアを探してきて
- テディベアはどこにあるか見つけて

[*1] SIGVerse, http://www.sigverse.org/wiki/jp/.
[*2] Sentence Generator 2011 for the General Purpose Service Robots, http://komeisugiura.jp/software/software_jp.html.

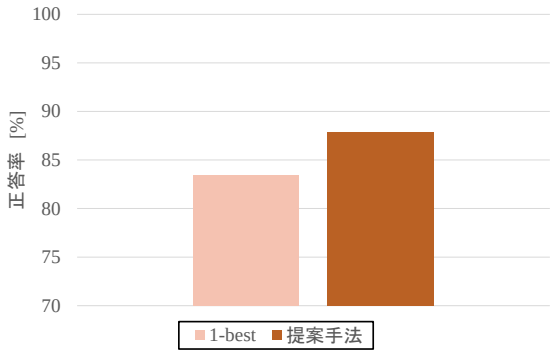

図 2　音声命令理解の精度

評価に用いる命令文はデータセットからランダムに 157 文を選択した．この選択した命令文を，ノイズがない静かな環境で読み上げ，その音声を録音した．提案モデルの学習に用いたデータは，残りの 1412 文である．ベースラインとして，音声認識・SVM・CRF の 1-best のみを利用した手法 (1-best) を用いた．

図 2 が音声命令理解の結果であり，157 文中の正答率を示している．この図より，提案手法のモデルはマルチモーダル情報を付与することで正答率を向上させることができた．

5 まとめ

本稿では，GMM を用いより実環境に近い条件においても対応できる音声命令理解モデルを提案した．提案モデルは，物体が移動してしまうような複雑な環境下においても正確に環境情報を取得することができ，正答率を向上させることができた．

今後，録音する音声にノイズを付与した条件で，提案モデルの性能を検証する．

参考文献

[1] 小堀嵩博, 中村友昭, 長井隆行, 岩橋直人, 船越孝太郎, 中野幹生, 金子正秀. 環境情報を考慮したロボットによる音声命令理解. In *The 30th Annual Conference of the Japanese Society for Artificial Intelligence*, 2016.

[2] Stefan Schiffer, Niklas Hoppe, and Gerhard Lakemeyer. Natural language interpretation for an interactive service robot in domestic domains. In *Agents and Artificial Intelligence*, pp. 39–53. Springer, 2012.

[3] Xiaoping Chen, Wei Shuai, Jiangchuan Liu, Song Liu, Ningyang Wang, Dongcai Lu, Yingfeng Chen, and Keke Tang. Kejia: The intelligent domestic robot for robocup@ home 2015, 2015. Teamdescription papers: RoboCup@Home League.

[4] Deb Roy and Niloy Mukherjee. Towards situated speech understanding: Visual context priming of language models. *Computer Speech & Language*, Vol. 19, No. 2, pp. 227–248, 2005.

[5] Hiroshi Murase and Vinod V Vasudevan. Fast visual search using focused color matching–active search. *Systems and Computers in Japan*, Vol. 31, No. 9, pp. 81–88, 2000.